Rhythm Pattern Representations for Tempo Detection in Music

Sankalp Gulati, Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay, Mumbai 400076, India
{sankalpg, prao}@ee.iitb.ac.in

ABSTRACT

Detection of perceived tempo of music is an important aspect of music information retrieval. Perceived tempo depends in a complex manner on the rhythm structure of the audio signal. Machine learning approaches, proposed recently, avoid peak picking and use rhythm pattern matching with stored tempo annotated songs in the database. We investigate different signal processing methods for rhythm pattern extraction and evaluate these for the music tempo detection task. We also investigate the effect of using additional information about the rhythmic style on the performance of the tempo detection system. The different systems are comparatively evaluated on a standard Ballroom Dance music database and an Indian music database.

1. INTRODUCTION

Tempo estimation of music has been an active area of research in the last decade due to its great importance in Music Information Retrieval (MIR) applications ranging from automatic cover-song identification, mood recognition and play-list generation to making seamless remixes by DJs. Depending upon the application, the term tempo has been variedly used to denote different periodicities in the audio, e.g. score tempo, foot-tapping tempo and perceived tempo to name a few. Perceptual tempo is the term used for the beats per minute (BPM) value, which is in accordance with the human perception of speed of the music (i.e. fast-slow categorization). Automatic tempo detectors are based on the detection of salient periodicity and are hence prone to octave errors (commonly doubling or halving of tempo). It is obvious that in the applications mentioned above, such octave errors can have a disastrous impact.

The process of parameterization of perceived speed of music into a single BPM value can, of course, suffer from genuine ambiguity. It is possible that listeners perceive different tempi in the same piece of audio, as shown in a tapping experiment [1]. But for practical purposes we can still move on by considering the ground-truth value to be the BPM value with which a majority of the listeners agree.

Figure 1 provides a generic framework used by most available tempo estimation algorithms. The front end module extracts a periodicity function or "rhythm pattern" representing the periodicities of events in the audio and their relative strengths. The events represent the instants of musical stress arising from musically salient events such as note changes and percussion transients captured in the accent signal. Various accent signal extraction methods are presented in [2]. The next step of periodicity analysis of the accent signal provides a rhythm pattern that could be in the form of an inter-onset interval histogram (IOIH) [3], autocorrelation function (ACF) [4] or comb filter output [5].

The rhythm pattern is then used in the subsequent module which extracts the salient periodicity from it. A comparison of the major tempo estimation algorithms can be found in [6]. Earlier algorithms for tempo estimation were designed to pick the prominent peaks from the rhythm pattern [5]. Considering the predilection of human beings towards certain tempo range, a biasing function around 120 BPM is advocated in [4]. However the task of perceptual tempo estimation requires more sophisticated approaches for picking up periodicities in accordance with the human perception of speed. There are various factors affecting our perception of speed and an exact understanding of the human perception of pulsation, periodicity and speed is yet not very clear. It is also acknowledged that there is a strong relation between rhythm pattern and tempo although the precise dependence is not known. In an effort to exploit this aspect given the limited understanding of the inter-relation, a machine-learning approach was proposed by Seyerlehner et.al. [9]. Rhythm pattern matching was applied to estimate the perceived tempo from a large manually labeled database of patterns based on the assumption that music with similar rhythmic structure have similar tempi. A further enhancement of this work involved incorporating k-NN regression for superior pattern matching [10].

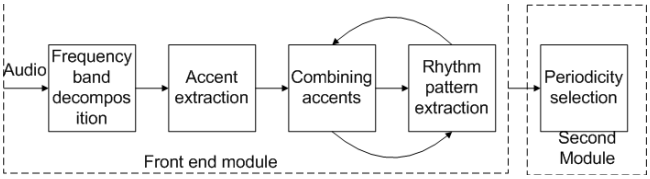

Figure 1. Generic framework for tempo estimation algorithms

Although very promising, the rhythm pattern matching based methods are also afflicted to an extent by the "moderate tempo tendency" [6], where for slow songs the estimated tempo is fast and for fast songs, slow. This ubiquitous problem of octave errors in all the tempo estimation algorithms motivates continuing

research on tempo estimation. Possible algorithmic issues for deeper investigation include the derivation of the rhythm pattern from the accent signal so that it captures better those aspects that are related to perceived tempo, and the similarity matching of rhythm patterns themselves [6]. The present work focuses on the first issue by way of designing a robust front end module that generates the required rhythm patterns. We explore different approaches to extract rhythm patterns from the perspective of improving the rhythmic similarity achieved and thus reducing the errors in system arising from ambiguous matches that lead to octave errors. The performances of three different approaches for rhythm pattern extraction are compared in terms of accuracy in perceived tempo estimation on a standard Ballroom Dance music database [11] and an Indian music database. Finally we explore the utility of additional information related to rhythmic style for the tempo detection task.

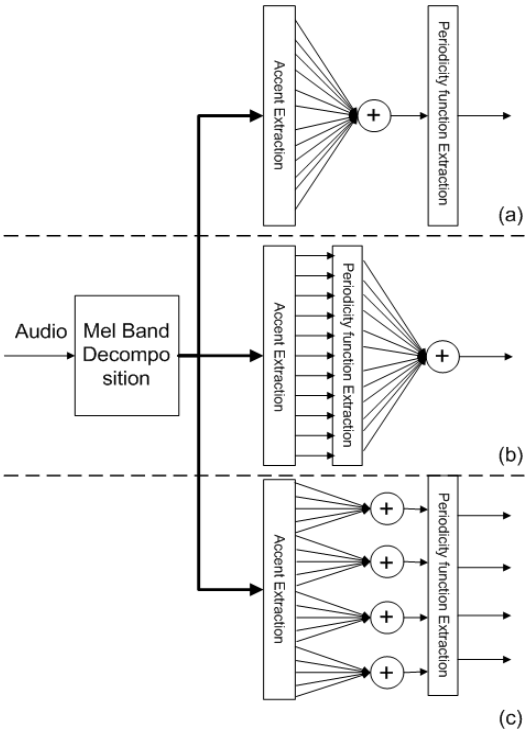

Figure 2. Block diagram of methods (a) reference (RM) (b) method 1 (M1) and (c) method 2 (M2)

2. METHOD

We use the basic approach of Seyerlehner et al. [9] in which the tempo estimation is modeled as two stage process of rhythm pattern extraction and tempo estimation by k-NN matching with stored rhythm patterns. The rhythm patterns pre-computed from the annotated songs in the database are used to detect the perceived tempo of the query song via k-NN matching of rhythm patterns. The method described in [9] is followed for the extraction of rhythm pattern with the exception that the audio bandwidth considered for the calculation of onset signal is increased from 4 to 8 kHz. Figure 2.a describes the block diagram for this reference method. A log magnitude 52 channel Mel spectrogram is computed with 32 ms window size and 4 ms hop between frames. For each of these Mel bands an onset signal is extracted by performing a first order difference of the filtered waveform followed by half wave rectification. The onset signals of all 52 bands are then additively combined after removing d.c. offset by high pass filtering. An ACF is computed from this combined onset signal up to 4 s lag. The ACF is smoothened by using an averaging filter of length 20 lag samples to obtain the rhythm pattern for the next stage of tempo estimation. This method of extracting rhythm pattern is taken as the reference method (RM) for comparison in subsequent sections.

For the estimation of perceived tempo using these rhythmic patterns, a k-nearest-neighbors (k=5) classifier is used to search for songs which have most similar rhythmic patterns. The tempo value most frequent among these neighbors is selected. The distance measure for similarity of rhythm patterns x and y is the normalized cross correlation as in Eq. 1.

$$r(x,y) = \frac{\sum_{j=1}^{N}(x_j - m_x)(y_j - m_y)}{\sqrt{\sum_{j=1}^{N}(x_j - m_x)^2 \sum_{j=1}^{N}(y_j - m_y)^2}} \quad (1)$$

Where, x_j is the j^{th} sample of x, m_x is the mean of x and N is the total number of samples in rhythm pattern.

Although the performance accuracy of this method is high, it suffers from the problem of octave errors as observed in preliminary testing. An observation of the octave errors revealed that they could be attributed to expressive percussion insertions in a rhythmic cycle and strong vocal onsets between beats, especially for low tempo songs. These events typically do not influence our perception of the speed. We term these as extra-metrical events. These events can cause the rhythm pattern to resemble that of database songs in a different tempo octave. In order to effectively capture and enhance only the aspects essential for rhythmic similarity, rhythm patterns should be relatively robust to extra-metrical events. We investigate two alternate ways of rhythm pattern representation with a view to improving the robustness to accent signal variations due to the extra-metrical events hoping to thus reduce the number of octave errors in tempo estimation. The methods differ from each other and from the original approach of [9] in the way the band-level periodicities are combined to form the rhythm pattern.

2.1 Method 1 for rhythm pattern extraction

In this method (M1) (depicted in Fig. 2(b)) the periodicity estimation is performed on the individual onset signals of different frequency bands, and later integrated to form a rhythm pattern. By doing so the distortion due to extra-metrical onsets is reduced because the undesired onsets present in a frequency band do not interact with the metrical onsets of other frequency bands. The biasing of periodicity function towards faster metrical levels, if periodicity estimation is performed after multiband onset integration is discussed in [6]. We try to reduce this biasing by performing periodicity estimation first. A rhythm pattern is desirable that corresponds to the main metrical events of a particular rhythmic style. This is because there can be numerous variations of a basic rhythm structure and to have songs with similar variation in training set is practically not feasible. An onset signal is extracted from each Mel band filtered signal as in the reference method. The ACF is computed for each band-level onset signal after removing d.c. offset by high pass filtering. The ACFs are added together and the resulting representation is smoothened.

2.2 Method 2 for rhythm pattern extraction

Figure 2(c) shows the block diagram for this method (M2) of periodicity representation. In this method instead of representing periodicities in audio by a single rhythm pattern, we propose using multiple rhythm patterns extracted from different frequency bands. The basic idea is to restrict the distortion due to extrametrical events to the same frequency band. While calculating distance measure for similarity between two songs based on multiband rhythm pattern, a criteria for band selection can be devised which bias the contribution of individual rhythm pattern in overall distance measure depending upon the amount of distortion it had undergone.

In this method, the onset signals of Mel filtered bands whose frequency fall under the frequency bands considered are added together. In the current work, band boundaries chosen are 0, 200, 1000, 4000, 8000 Hz forming four bands B1, B2, B3 and B4 respectively. For each of the band onset signals, a smoothened ACF is computed up to 4s to get the full multiband rhythm pattern that is used for rhythmic similarity matching.

The distance measure for multiband rhythm pattern representation is calculated according to Eq. 2.

$$d = \sum_{i=1}^{M} a_i * r(\overrightarrow{x_i}, \overrightarrow{y_i}) \qquad (2)$$

Where, $\overrightarrow{x_i}$ and $\overrightarrow{y_i}$ are rhythm patterns of the i^{th} band, M is the total number of frequency bands, a_i is the weight for distance measure of the i^{th} band, and $r(x,y)$ is given by Eq.1. Various choices for the coefficients a_i are considered later to emphasize certain bands over the rest.

3. INCORPORATING RHYTHM STYLE

The effect of knowledge of rhythm style in the tempo estimation task has been studied in [7,8]. A style dependent weighting function is applied to the periodicity function using the statistical knowledge of the tempo ranges for a particular style. However incorporating knowledge of rhythmic style without using any statistical information about the tempo ranges would be more useful especially when tempo ranges of different rhythmic styles bear a heavy overlap. We utilize this information by restricting the search range of rhythm pattern of query song to the patterns of the same rhythmic style, known through ground truth labels. Also, as the musical meter of a song imparts partial information about the rhythmic style, it is interesting to analyze the effect of using only meter information on tempo estimation. In order to fully automate this task we use the method described in [12] for meter detection.

4. EXPERIMENTS AND EVALUATION

The methods discussed in previous sections form three distinct front end modules to the tempo estimation block, providing different rhythm pattern representations. Each of the methods is integrated with the k-NN classifier as described in Section 2. Regression is used in the pattern matching block as described in [10] to compensate for the relatively small size of the Indian Music database. All the three systems are evaluated for different cases i.e. with and without using rhythm style information, using meter information and using estimated meter information.

4.1 Database

The evaluation of the different methods is done on two databases (DB). The first is the well-known ballroom dance database (BDDB) containing 698 30 sec duration audio clips, widely used in the tempo detection evaluations [11]. It comprises of 8 different ballroom dance styles (Jive, Quickstep, Tango, Waltz, Viennese Waltz, Samba, Cha cha cha and Rumba). This database is available annotated with BPM values. No cross checking of tempo values was done by the authors.

The second database is an Indian Music Database (IMDB) constructed by the authors comprising 175 30 sec duration audio clips belonging to three different rhythm styles of Hindustani Classical Music (Dadra (66), Kaherwa (66) and Bhajan Theka (43)). The Dadra style is in triple meter whereas other two follow double meter. The tempo annotation was done manually by the authors so that the BPM values are consistent with the perceived speed of the song.

4.2 Evaluation

All the methods are evaluated by "Leave one out" cross-validation, i.e. each song in the database is used as a test token with the remaining songs comprising the training set (from which nearest neighbors are selected). A tempo estimate is marked correct only if it falls within 4% of the annotated tempo value. We evaluate the meter detection module separately on both the databases based on the ground-truth meter for the particular rhythm style.

In order to get an insight into the rhythm patterns of the individual bands considered in M2, we evaluate the system using rhythm patterns extracted from only one band at a time. Based on the performance of individual bands we choose various combinations of the coefficients a_i to bias the contribution of bands giving high accuracy and select the best performing combination for further evaluation and comparison of M2.

Table 1. Performance (%) of tempo detection using rhythm pattern extracted from individual bands (B1-B4) in M2.

DB	B1	B2	B3	B4
BDDB	87.1	85.4	83.1	84.5
IMDB	84	81.1	81.7	75.4

Table 2. Performance (%) of the 3 rhythm pattern extraction methods (M1, M2, and RM) on two databases for different cases of side information: rhythm style (RS), meter (MTR1) and estimated meter (MTR2); MV is majority voting and NN (% incorrect) by all methods

DB	INFO	RM	M1	M2	MV	NN
BDDB	-	87.10	87.54	87.96	88.25	8.6
BDDB	RS	92.12	92.26	92.69	92.4	6.73
BDDB	MTR1	88.10	89.54	89.25	89.4	8.16
BDDB	MTR2	87.39	87.96	88.96	88.83	8.16
IMDB	-	89.71	92	92	92	3.43
IMDB	RS	94.29	94.29	94.86	94.86	2.29
IMDB	MTR1	93.71	93.71	94.29	93.71	2.29
IMDB	MTR2	90.29	92.57	92	92.57	3.43

5. RESULTS AND DISCUSSION

Table 1 summarizes the performance of individual bands in M2 for tempo estimation. The performance variations across bands may be attributed to the various amounts of perturbation of the rhythm pattern of the band due to extra-metrical events manifested in that band. We see that the 1st band performs best (B1); it captures only the bass components within 200 Hz. This is followed by 2nd band (B2). We found that combining distance measure of these two bands (i.e. in Eq. 2, set a1=a2=1, and a3=a4=0) resulted in the overall highest tempo detection accuracy for M2.

The performance accuracy of all three methods is summarized in Table 2. We notice that the accuracy is quite high for all the methods. We see that M1 has provided slight improvement over reference method (RM) which is further improved upon by M2. We observe that using rhythm style information even without including any statistical information about the tempo ranges improves the result to a good extent (approximately by 4%). The information about the meter of the song alone improves the performance, especially for IMDB where meter classifies the rhythm styles to a good extent. The accuracy obtained in meter estimation itself was observed to be 89.83% for BDDB and 88% for IMDB. However due to the remaining 10-12% errors the performance accuracy of the system using estimated meter is reduced. NN is the percentage of songs incorrectly rated by all three methods. An analysis of these songs revealed that several of these correspond to tempo outliers considering songs of same rhythm style. Hence the rhythm pattern matching is poor. Further, in many cases the extra-metrical onsets are too strong to be suppressed by any method. The number of errors made by any single method and that by all methods indicates that different rhythm pattern representations contain complementary information that can be potentially exploited by a suitable combination approach. However we see from Table 2 that majority voting (MV) is not effective enough and other combination classifiers need to be investigated.

6. CONCLUSION

Rhythm pattern matching based tempo detection was found to perform well on the test databases. Octave errors were observed to arise mainly due to the presence of extra-metrical onsets in the audio signal. Performing periodicity analysis before combining accent signals was found to help in reducing the effect of irregular events on the rhythm pattern. Further, using only selected low frequency bands to derive the accent signal improves performance. It is interesting to note that knowledge of rhythm style even without using style dependent tempo ranges can serve to improve system accuracy for the tempo estimation task. Further, meter information alone can help in reducing octave errors in the rhythm pattern matching approach to tempo estimation. Future work should target improving automatic meter detection, and testing on other, larger music databases.

7. REFERENCES

[1] McKinney, M.F., Moelants, D. 2004. Tempo Perception and Musical Content: What makes a piece fast, slow or temporally ambiguous? In *Proceedings of the International Conference on Music Perception and Cognition (ICMPC'04)*, Evanston, USA.

[2] Dixon, S. 2006. Onset Detection Revisited. In *Proceedings of the International Conference on Digital Audio Effects (DAFx'06)*, Montreal, Canada.

[3] Dixon, S. 2001. Automatic extraction of tempo and beat from expressive performances, *J. New Music Res.*, vol. 30, no. 1, pp. 39–58.

[4] Ellis, D. P. 2007. Beat tracking by dynamic programming. *J. New Music Res.*, vol. 36, no. 1, pp. 51–60.

[5] Scheirer, E. 1998. Tempo and beat analysis of acoustic musical signals. *J.Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601.

[6] Gouyon, F.,Klapuri, A., Dixon, S.,Alonso, M.,Tzanetakis, G., Uhle, C. and Cano, P. 2006. An experimental comparison of audio tempo induction algorithms. *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1832–1844.

[7] Davies, M. E. P., Plumbley, M. D. 2008. Exploring the effect of rhythmic style classification on automatic tempo estimation, In *Proceedings European Signal Processing Conf.*

[8] Schuller, B., Eyben, F., Rigoll. G. 2008. Tango or Waltz?: Putting ballroom dance style into tempo detection. E*URASIP Journal on Audio, Speech, and Music Processing*, 2008 (846135):12 pages, 2008.

[9] Seyerlehner, K., Widmer, G., Schnitzer, D. 2007. From rhythm patterns to perceived tempo. In *Proceedings. 8th Int. Conf. Music Inf. Retrieval (ISMIR- 07)*, Vienna, Austria

[10] Eronen, A. J., Klapuri. A. 2010. Music tempo estimation with k-NN regression, In *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, January

[11] Gouyon, F., Dixon, S., Pampalk, E., Widmer, G. 2004. Evaluating rhythmic descriptors for musical genre classification, *In Proceedings. AES 25th Int. Conf.*, New York, 2004, pp. 196–204.

[12] Schuller, B., Eyben, F. and Rigoll, G. 2007. Fast and robust meter and tempo recognition for the automatic discrimination of ballroom dance styles, *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP '07), pp. 217–220, Honolulu, Hawaii, USA, April.